VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

Fakulta elektrotechniky
a komunikačních technologií

BAKALÁŘSKÁ PRÁCE

Brno, 2021 David Slavíček

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA ELEKTROTECHNIKY A KOMUNIKAČNÍCH TECHNOLOGIÍ
FACULTY OF ELECTRICAL ENGINEERING AND COMMUNICATION

ÚSTAV BIOMEDICÍNSKÉHO INŽENÝRSTVÍ
DEPARTMENT OF BIOMEDICAL ENGINEERING

VLIV SEKVENÁTORŮ DRUHÉ A TŘETÍ GENERACE NA CGMLST ANALÝZU BLÍZCE PŘÍBUZNÝCH BAKTERIÁLNÍCH KMENŮ.

INFLUENCE OF SECOND AND THIRD GENERATION SEQUENCERS ON CGMLST ANALYSIS OF CLOSELY-RELATED BACTERIAL STRAINS.

BAKALÁŘSKÁ PRÁCE
BACHELOR'S THESIS

AUTOR PRÁCE — **David Slavíček**
AUTHOR

VEDOUCÍ PRÁCE — **Ing. Markéta Nykrýnová**
SUPERVISOR

BRNO 2021

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
FAKULTA ELEKTROTECHNIKY A KOMUNIKAČNÍCH TECHNOLOGIÍ

Bakalářská práce

bakalářský studijní program **Biomedicínská technika a bioinformatika**

Ústav biomedicínského inženýrství

Student: David Slavíček ***ID:*** 211212
Ročník: 3 ***Akademický rok:*** 2020/21

NÁZEV TÉMATU:

Vliv sekvenátorů druhé a třetí generace na cgMLST analýzu blízce příbuzných bakteriálních kmenů.

POKYNY PRO VYPRACOVÁNÍ:

1) Vypracujte literární rešerši na téma sekvenační technologie druhé a třetí generace a zaměřte se i na následné sestavování genomů. 2) Navrhněte metodu pro sestavení dat z druhé a třetí generace sekvenátorů a na základě navržené metody sestavte genomy poskytnuté z FN Brno. 3) Navrhněte algoritmus pro nalezení a porovnání genů, které se využívají pro cgMLST a dílčí části realizujte. 4) Navržený algoritmus implementujte ve vhodném programovacím prostředí a otestujte na sestavených genomech. 5) Proveďte klasifikaci genomů na základě cgMLST. 6) Výsledky vhodně graficky zobrazte a diskutujte.

DOPORUČENÁ LITERATURA:

[1] SHENDURE, Jay, Shankar BALASUBRAMANIAN, George M. CHURCH, Walter GILBERT, Jane ROGERS, Jeffery A. SCHLOSS a Robert H. WATERSTON. DNA sequencing at 40: past, present and future. Nature. 2017, 550(7676), 345-353. DOI:10.1038/nature24286

[2] SCHÜRCH, Anita C., Sergio ARREDONDO-ALONSO, Rob J.L. WILLEMS a Richard V. GOERING. Whole genome sequencing options for bacterial strain typing and epidemiologic analysis based on single nucleotide polymorphism versus gene-by-gene–based approaches. Clinical Microbiology and Infection. 2018, 24(4), 350-354. DOI:10.1016/j.cmi.2017.12.016

Termín zadání: 8.2.2021 ***Termín odevzdání:*** 28.5.2021

Vedoucí práce: Ing. Markéta Nykrýnová

doc. Ing. Jana Kolářová, Ph.D.
předseda rady studijního programu

ABSTRAKT

Tato bakalářská práce se zabývá vlivem sekvenátorů druhé a třetí generace na cgMLST analýzu bakteriálního genomu. V teoretické části byly popsány některé sekvenátory druhé a třetí generace a principy sestavování genomu. V praktické části pak byly použity nasekvenované genomy bakterií z FN Brno. Jedná se o genomy šesti bakterií *Klebsiella pneumoniae*, které byly nasekvenovány na dvou různých sekvenátorech, Illumina Miseq a Oxford Nanopore Technologies Minion. Tyto genomy byly sestaveny. Pro cgMLST analýzu byly vybrány vhodné geny a vyřazeny genomy, které se nepodařilo sestavit v dostatečné kvalitě. Následně byla cgMLST analýza provedena a výsledky graficky zobrazeny.

KLÍČOVÁ SLOVA

cgMLST, sekvenace, Illumina, Oxford Nanopore Technologies, sestavení genomu, SPAdes

ABSTRACT

This bachelor thesis is about influence of sequncers of second and third generation on cgMLST analysis of bacterial genome. In the theoretical section were described selected sequencers of second and third generation and genome assembly principles. In the practical part were assembled six genomes of *Klebsiella pneumoniae* bacteria from University Hospital Brno. The genomes were sequenced each on two different sequencers, Illumina Miseq and Oxford Nanopore Technologies Minion. The genomes were assembled. Suitable genes were selected and insufficient quality genomes removed for the cgMLST analysis. The cgMLST analysis was performed and the results are shown in graphs.

KEYWORDS

cgMLST, sequencing, Illumina, Oxford Nanopore Technologies, genome assembly, SPAdes

SLAVÍČEK, David. *Vliv sekvenátorů druhé a třetí generace na cgMLST analýzu blízce příbuzných bakteriálních kmenů*. Brno, 2021, 44 s. Bakalářská práce. Vysoké učení technické v Brně, Fakulta elektrotechniky a komunikačních technologií, Ústav biomedicínského inženýrství. Vedoucí práce: Ing. Markéta Nykrýnová

PROHLÁŠENÍ

Prohlašuji, že svou bakalářskou práci na téma „Vliv sekvenátorů druhé a třetí generace na cgMLST analýzu blízce příbuzných bakteriálních kmenů“ jsem vypracoval samostatně pod vedením vedoucího bakalářské práce a s použitím odborné literatury a dalších informačních zdrojů, které jsou všechny citovány v práci a uvedeny v seznamu literatury na konci práce.

Jako autor uvedené bakalářské práce dále prohlašuji, že v souvislosti s vytvořením této bakalářské práce jsem neporušil autorská práva třetích osob, zejména jsem nezasáhl nedovoleným způsobem do cizích autorských práv osobnostních a/nebo majetkových a jsem si plně vědom následků porušení ustanovení § 11 a následujících autorského zákona č. 121/2000 Sb., o právu autorském, o právech souvisejících s právem autorským a o změně některých zákonů (autorský zákon), ve znění pozdějších předpisů, včetně možných trestněprávních důsledků vyplývajících z ustanovení části druhé, hlavy VI. díl 4 Trestního zákoníku č. 40/2009 Sb.

Brno ....

podpis autora

PODĚKOVÁNÍ

Rád bych poděkoval vedoucí diplomové práce paní Ing. Markétě Nykrýnové za odborné vedení, konzultace, trpělivost a podnětné návrhy k práci.

Výpočetní zdroje byly poskytnuty z projektu "e-Infrastruktura CZ" (e-INFRA LM2018140) řešeného v rámci programu Projects of Large Research, Development and Innovations Infrastructures.

Obsah

Seznam obrázků

Úvod

V posledních letech výrazně klesá cena sekvenace, tedy čtení sekvence DNA, RNA, nebo proteinu. S klesající cenou se otevírá mnoho různých možností využití sekvenace. Jednou z nich je kontrola šíření nakažlivých nemocí, kterou lze provádět například pomocí cgMLST analýzy. Tato bakalářská práce spočívá ve vytoření literární rešerše na téma sekvenační technologie druhé a třetí generace a sestavení šesti genomů bakterie *Klebsiella pneumoniae* nasekvenovaných na dvou různých platformách, provedení cgMLST analýzy a grafickém zobrazení výsledků. V první kapitole je blíže popsán proces sekvenace. Zmíněny jsou sekvenační přístroje druhé a třetí generace. V druhé kapitole je popsáno, jak lze z dat ze sekvenačních přístrojů sestavovat celé sekvence a popsány, některé programy k tomu určené. Ve třetí kapitole je popsána typizace. Čtvrtá kapitola se zabývá shlukovacími metodami, vhodnými ke grafickému znázornění výsledků typizace. V páté kapitole je popsáno provedení samotné cgMLST analýzy. Šestá kapitola shrnuje dosažené výsledky. [1]

1 Sekvenace

1.1 Druhá generace sekvenování

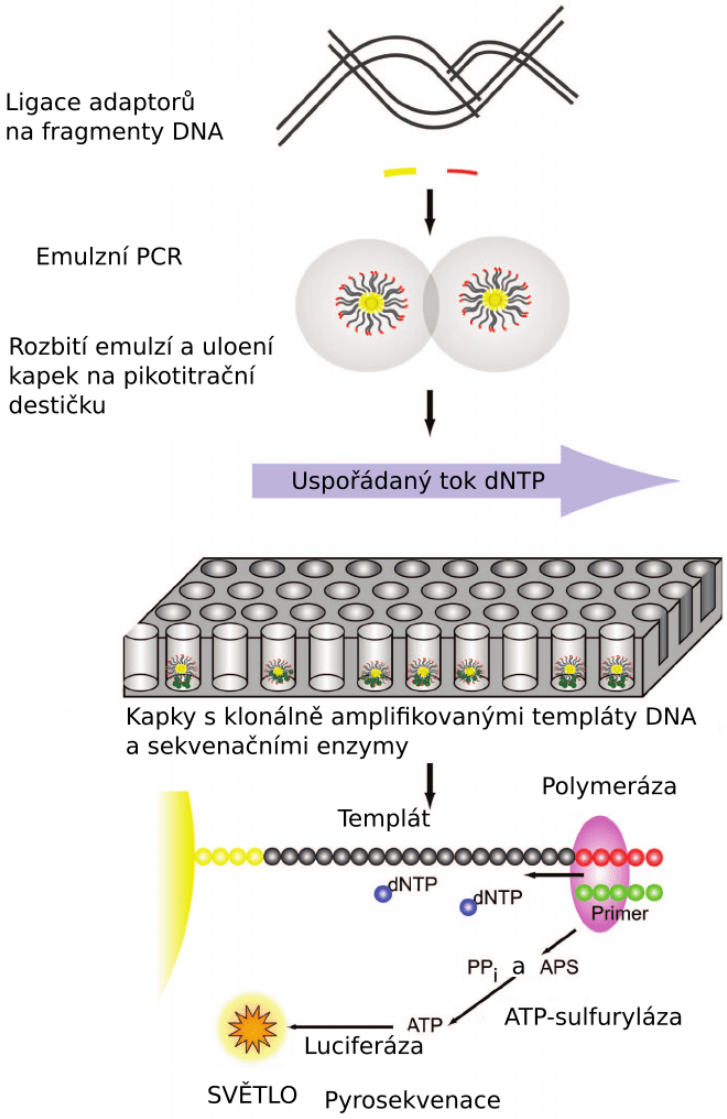
Sekvenační přístroje druhé generace (z angl. next-generation) přináší proti sekvenačním přístrojům první generace dvě zásadní výhody. Těmi jsou výrazné zlevnění a zrychlení sekvenace. Výhod je dosaženo masivní paralelizací sekvenace. Obecně poskytují sekvenátory druhé generace poměrně krátká čtení. [2]

1.1.1 Roche 454

Sekvenátory založené na technologii 454 byly první sekvenátory druhé generace. Sekvenátory využívají pikotitrační destičku s jamkami. Knihovna DNA je připravována fragmentací na fragmenty o délkách několik set bazí. Na konce fragmentů jsou ligovány adaptory. Na pikotitrační destičce jsou přichystány kapičky vody v oleji. Kapičky obsahují na povrchu templáty DNA komplementární k adaptorům, namnožené pomocí polymerázové řetězové reakce (PCR). Postupně je na pikotitrační destičku přidán vždy jeden deoxynukleotrifosfát (dNTP), poté je vytvořen snímek destičky pomocí CCD čipu. Následně je původní dNTP vymyt a proces opakován s dalším dNTP. Světlo zaznamenávané CCD čipem emituje enzym luciferáza, při kontaktu s pyrofosfátem. Pyrofosfát je uvolňován při připojení dNTP do DNA. Enzym luciferáza je přítomný v jamkách na pikotitrační destičce. Princip fungování tohoto typu sekvenátorů je znázorněn na obrázku 1.1. [3, 4]

1.1.2 Illumina

Firma Illumina využívá reverzibilních terminátorů. Dvouvláknová DNA je nejprve naštěpena, následně jsou na její konce ligovány adaptory a je denaturována na jednovláknovou DNA. Jednovláknová DNA je navázána na skleněnou destičku s přichycenými úseky DNA komplementárními k adaptorům. Poté je provedena PCR, čímž jsou vytvořeny shluky stejných úseků jednovláknové DNA. Reverzní vlákna jsou následně odstraněna. Proces pokračuje čtením syntézou. V každém kroku je přidána polymeráza a všechny čtyři deoxynukleotidy s reverzibilním terminátorem sloužícím zároveň jako barevné značení. Po excitaci laserem je detekováno emitované světlo, kde každý z deoxynukleotidů je značen jinou vlnovou délkou. Terminátory jsou odstraněny a následuje další cyklus. Přístroje této firmy jsou schopny osekvenovat maximálně 300 bazí z každé strany fragmentu nukleové kyseliny. Výhodou této platformy je vysoká přesnost. Kroky běhu sekvenátorů od firmy Illumina jsou znázorněny na obrázku 1.2. [6, 7]

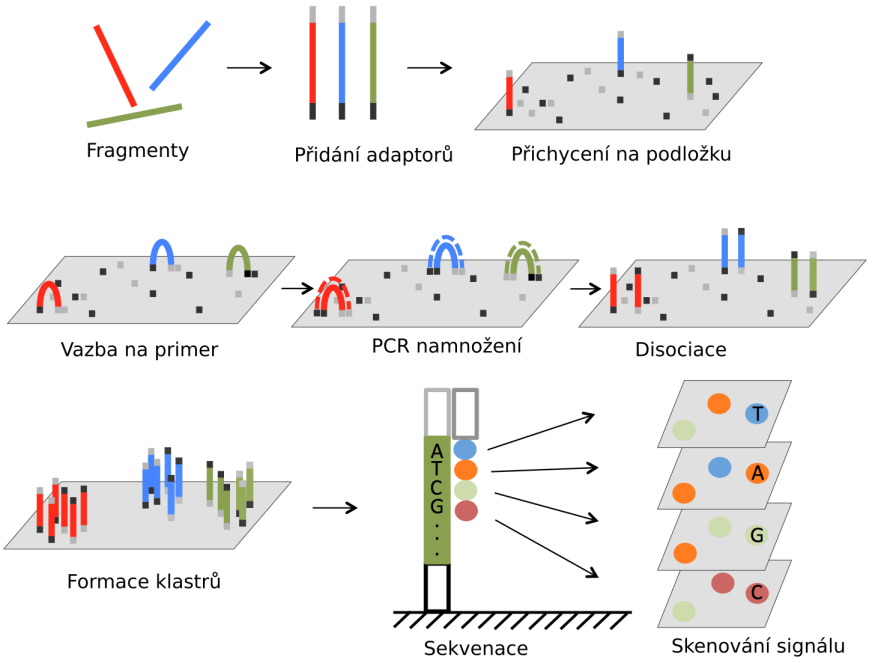


Obr. 1.1: Princip fungování sekvenátorů Roche 454. Převzato z [5]

1.2 Sekvenační přístroje třetí generace

1.2.1 PacBio

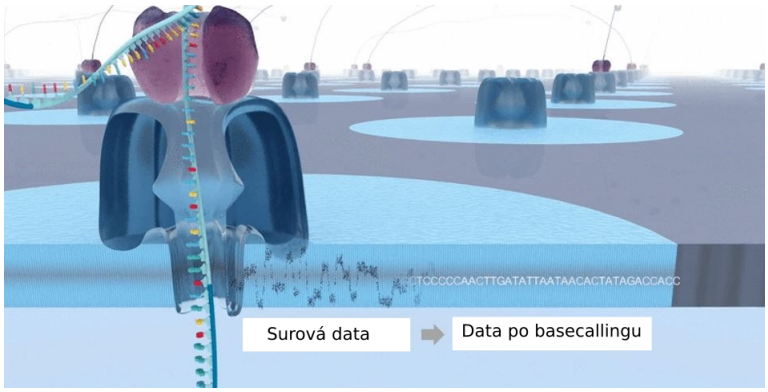
Sekvenátor firmy Pacific Biosciences produkuje čtení o průměrné délce přes 2500 bp a delší čtení mohou dosahovat délek přes 10 000 bp. Nevýhodou je až 15 % chybovost u dlouhých čtení. Chybovost může být výrazně snížena metodou "circular consensus sequencing". [9, 10]



Obr. 1.2: Princip fungování sekvenátorů firmy Illumina. Převzato z [8]

1.2.2 Oxford Nanopore

Firma Oxford Nanopore Technologies využívá měření změny velikosti elektrického proudu protékajícího nanopórem, když je nanopórem protahována DNA. Sekvenovací jednotka sestává z nevodivé membrány, ve které jsou umístěny nanopóry. Na membráně je elektrické napětí a zařízení měří proud protékající z jedné strany na druhou. Na nanopór nasedá protein s uchycenou DNA. Protein zajišťuje přenos DNA k nanopóru a zároveň reguluje rychlost průchodu DNA nanopórem. Výstupem je velikost elektrického proudu v čase, ze které je následně dopočítáno pořadí bazí. Velikost elektrického proudu v čase je ukládána v souborech s příponou .fast5. Sekvenátor má 512 kanálů a může tedy sekvenovat až 512 molekul DNA současně. V každém kanálu jsou čtyři nanopóry, DNA však vždy prochází jen jedním. Délka čtení je maximálně několik set tisíc párů bazí, tato výhoda je vykoupena poměrně nízkou přesností v rozmezí 65 - 88 %. Průběh sekvenace tímto sekvenátorem je znázorněn na obrázku 1.3. [11, 12]



Obr. 1.3: Nanopór s nasednutým proteinem a procházející nukleovou kyselinou. Převzato z [18]

2 Sestavování genomu

Sestavování genomu (z angl. genome assembly) je proces, při kterém skládáním čtení ze sekvenátoru získáváme buď sekvenci celého genomu, nebo několik delších sekvencí, nazývaných kontigy. U kontigů nejsme schopni určit, v jakém pořadí se za sebou vyskytují a jak velké mezery mezi nimi jsou. Víme jen, že se někde v genomu nacházejí. Sestavování genomu můžeme provádět dvěma základními způsoby, a to *de novo* nebo k mapováním k referenční sekvenci. Výhodou skládání *de novo* je, že nepotřebujeme referenční sekvenci, nevýhodou jsou komplikace při dlouhých repetitivních úsecích, zejména pokud použijeme sekvenátor druhé generace, který produkuje krátká čtení. Další možností je použití hybridního sestavování, tedy využití dat z různých generací sekvenátorů. [2]

2.1 Eliminace špatných čtení počítáním k-merů

Sekvenátory druhé generace dosahují poměrně vysoké přesnosti čtení, i přesto potřebujeme eliminovat chybně přečtené báze, aby bylo možné složit genom. Jednoduchým způsobem detekování chybných čtení je počítání k-merů.

K-mer je podřetězec sekvence DNA nebo RNA o délce k. Sekvence ACGT má tedy tři 2-mery AC, CG, GT, dva 3-mery ACG a CGT a jeden 4-mer ACGT.

Nejprve si stanovíme k, které chceme pro eliminaci použít. Následně vytvoříme tabulku četnosti jednotlivých k-merů. Každý fragment přečtený sekvenátorem rozložíme na k-mery. Následně vezmeme každý k-mer, není-li v tabulce četnosti k-merů, přidáme jej do ní, v opačném případě inkrementujeme jeho četnost v tabulce. Když máme tabulku sestavenou, stanovíme si minimální přípustnou četnost k-meru. Fragmenty obsahující k-mery s menší než minimální přípustnou četností byly velmi pravděpodobně chybně přečteny sekvenátorem. [2]

2.2 De Bruijnovy grafy

De Bruijnovy grafy jsou orientované. Pokud sestavujeme sekvence pomocí de Bruijnových grafů, využijeme nejprve rozložení fragmentů na k-mery. Potom pro každý k-mer vytvoříme jeden vrchol, bez ohledu na celkový počet k-merů v souboru fragmentů. Pokud prvních k-1 znaků na konci jednoho k-meru je shodných s k-1 znaky na začátku druhého k-meru, vedeme hranu z vrcholu odpovídajícímu prvnímu k-meru do vrcholu odpovídajícímu druhému k-meru. Při hledání výsledné sekvence hledáme takovou cestu, abychom prošli v kuse každé čtení. [13, 14]

2.3 OLC

Metoda Overlap layout consensus sestává ze tří kroků. Prvním je vytvoření grafu překryvu čtení, druhým vyhledání kontigů v grafu a třetím určení sekvencí kontigů. K sestavení grafu překryvu čtení může být využit například suffixový strom nebo dynamické programování. Graf překryvů je orientovaný a jeho hranám jsou přiřazeny váhy. Vrcholy grafu překryvů jsou jednotlivá čtení. Pokud se dvě čtení překrývají v délce delší, než je daný požadovaný limit, jsou propojeny hranou, jejíž váha odpovídá délce překrytí. Po sestavení grafu jsou odstraňovány nepotřebné hrany. Nepotřebné hrany jsou ty, které lze dovodit z tranzitivity velkých překryvů. Dovození funguje tak, že pokud čtení A překrývá čtení B překryvem délky x, čtení B překrývá čtení C překryvem délky y a zároveň délka b čtení B je menší než x + y, pak čtení A překrývá čtení B s překryvem délky alespoň x + y - b. Po odstranění nepotřebných hran vzniknou v grafu lineární části, ze kterých jsou vytvořeny kontigy. Kontigy jsou vytvořeny konsensem. [15]

2.4 Assemblery

2.4.1 SPAdes

St. Petersburg genome assembler, neboli SPAdes, je assembler pro *de novo* sestavování využívající de Bruijnových grafů. Umožňuje sestavení genomů sekvenovaných na sekvenátorech druhé generace nebo hybridní skládání, tedy s využitím dat ze sekvenátorů druhé i třetí generace současně. [14, 16]

2.4.2 Burrows Wheeler aligner

Burrows Wheeler Aligner (BWA) je balík softwaru určený k zarovnávání sekvencí k referenčnímu genomu. Sestává ze tří algoritmů - BWA-backtrack, BWA-SW a BWA-MEM. Algoritmus BWA-backtrack je určen k zarovnávání sekvencí o délce do 100 bp. Zbylé dva algoritmy jsou určeny k zarovnávání bazí o délce 70 bp až 1 Mbp. V této práci byl využit alogoritmus BWA-MEM, který je oproti BWA-SW novější a je doporučen pro skládání dat z platformy Illumina. Výstupem je soubor s příponou sam. [17]

2.4.3 Guppy

Guppy je toolkit vydaný Oxford Nanopore Technologies, který slouží k basecallingu a dalšímu zpracování dat ze sekvenátorů této firmy. Toolkit Guppy lze využit

kromě basecallingu také k demultiplexaci a skládání k referenci. Pro basecalling využívá Guppy rekurentní neuronové sítě. [18]

2.5 Hodnocení kvality sestavení

V tomto odstavci jsou čísla v rozmezí 0 až 100 označena písmenem x. Velmi používaným parametrem je Nx (např. N50) výpočet lze provést tak, že jsou seřazeny kontigy od nejdelšího po nejkratší a zapsány je za sebe. Poté je z výsledného řetězce ponecháno jen prvních x procent. Délka nejkratšího kontigu obsaženého v tomto začátku řetězce je parametr Nx. Dalším parametrem je NGx. K výpočtu tohoto parametru je třeba znát referenční genom. Parametr lze vypočíst obdobně jako Nx popsaný výše. Jediný rozdíl ve výpočtu je, že z řetězce seřazených kontigů není ponecháno prvních x procent délky řetězce, ale prvních x procent délky referenčního genomu. Při znalosti referenčního genomu je možné spočítat počet indelů na 100kb oproti referenci. Lze také porovnávat délku nejdelších kontigů, kde platí, že delší kontig je lepší. Je možné spočítat kontigy, které nejde vhodně zarovnat k referenci, kde je požadováno co nejméně takovýchto kontigů. Rovněž je možné počítat kontigy, které lze v referenci zarovnat na několik míst stejně dobře, kde je opět žádoucí co nejmenší počet. Pokud jsou při sestavování známy geny, u kterých lze očekávat, že budou v genomu obsaženy, lze počítat počet genů, které se v sestavení nachází. [26]

Jedním z programů složících k hodnocení kvality sestavení je QUAST. Dokáže vypočítat mimo jiné parametry popsané výše. Program také dokáže vykreslit různé grafy například graf parametru Nx, popsaného výše, kde na vodorovné ose je hodnota x a na svislé Nx. QUAST má i webové rozhraní. [26]

3 Typizace

Typizací rozumíme charakterizaci izolátů na nižší úrovni, než je rozlišení na druhy nebo poddruhy. Typizovat bakterie je vhodné abychom mohli monitorovat šíření nemocí a díky tomu šíření snadněji bránit. Metody typizace můžeme rozdělit do tří hlavních skupin. Těmi jsou metody založené na PCR, metody založené na analýze fragmentů DNA a metody založené na celogenomovém sekvenování (WGS). [1]

V současné době se díky rychlému vývoji sekvenátorů a poklesu ceny sekvenace a rozvoji bioinformatického softwaru do popředí dostávají metody založené na WGS. Pro typizaci bakterií jsou nejčastěji využívány metoda core genome multilokusová sekvenční typizace (cgMLST, z angl. core genome multilocus sequence typing) a metoda jednonukleotidových variací (SNV z angl. single nucleotide variant). [1, 19]

Metoda cgMLST spočívá ve vyhledání referenčních genů, a očíslování alel genů. Příbuznost organismů je pak dána počtem stejných nebo různých genů. Seznamy referenčních genů jsou volně dostupné v internetových databázích. Tato metoda se vyznačuje vysokou přesností a velmi dobrou přenositelností výsledků mezi laboratořemi. [1]

Metoda SNV je založena na mapování čtení nebo sestavených kontigů k referenční sekvenci. Tato metoda dosthuje ještě vyšší senzittivity než cgMLST. Data získaná metodou SNV, jsou však výrazně hůře porovnatelná mezi laboratořemi než u cgMLST. Proto je vhodná zejména pro malé analýzy. [1, 19]

4 Shlukovací metody

Shlukovací metody slouží k vyhledávání skupin, tedy shluků podobných objektů. Můžeme je dělit na hierarchické a nehierarchické. Hierarchické shlukovací metody vytváří nové shluky ze shluků dříve vytvořených, nehierarchické shlukovací metody vytváří všechny shluky současně. Významnou skupinou hierarchických shlukovacích metod jsou Aglomerativní shlukovací metody, na které je tento popis shlukovacích metod zaměřen. [33, 34, 35]

4.1 Aglomerativní hierarchické shlukovací metody

Aglomerativní hierarchické shlukovací metody se vyznačují tím, že v prvním kroku algoritmu má každý vstupní prvek svůj vlastní shluk. Tyto shluky jsou pak spojovány. [33]

Aglomerativní hierarchické shlukovací metody popsané v této práci jsou si vzájemně velmi podobné, liší se pouze ve způsobu, kterým jsou v každém kroku přepočteny vzdálenosti mezi shluky. Proto je v následujících dvou odstavcích popsána jejich společná sturktura a dále v textu přepočty vzdáleností shluků pro jednotlivé algoritmy. [33]

4.1.1 Společná struktura

Vstupem algoritmu je matice vzdáleností mezi prvky. V této bakalářské práci jsou prvky genomy. Algoritmus využívá matice vzdáleností mezi shluky, kterou průběžně aktualizuje. Na začátku obsahuje každý shluk právě jeden prvek a matice vzdáleností shluků je shodná s maticí vzdáleností prvků. [33, 34]

V každém kroku algoritmu jsou vybrány dva shluky s nejmenší vzájemnou vzdáleností a sloučeny v jeden shluk. Do matice s výsledky je zapsáno, které dva shluky byly v tomto kroku spojeny a jejich vzájemná vzdálenost. Následně je přepočtena matice vzdáleností shluků. Vzájemné vzdálenosti shluků které nebyly v tomto kroku slučovány zůstávají zachovány, nově vypočteny jsou pouze vzdálenosti nově sloučeného shluku se všemi ponechanými shluky. Krok je opakován dokud nejsou všechny prvky v jednom společném shluku. [33, 34]

4.1.2 UPGMA

Zkratka UPGMA znamená unweighted pair group method with arithmetic mean, tedy nevážená párová shlukovací metoda využívající aritmetický průměr. V algo-

ritmu UPGMA je nová vzdálenost dvou shluků A a B definována jako nevážený aritmetický průměr všech vzdáleností dvojic prvků a a b takových, že genom a náleží do shluku A a prvek b náleží do shluku B. [33, 36]

Při slučování větších shluků je výpočetně neefektivní počítat průměr mnoha vzájemných vzdáleností prvků. Uvažme shluk A vzniklý spojením shluků X, Y, u kterého počítáme jeho vzdálenost od shluku B. Stejného výsledku jako průměrováním vzdáleností prvků dle předchozího odstavce můžeme dosáhnout také tak, že vezmeme vzdálenost shluků X a B a vzdálenost shluků Y a B a uděláme jejich průměr vážený velikostmi shluků X a Y. Takto počítá shluky například funkce linkage ze softwarového balíku scipy, která byla využita v této práci. Výpočet vzdálenosti je popsán vzorcem níže. Počet prvků shluku Z v něm zančíme |Z|. [33]

$$d(A, B) = \frac{d(X, B) \cdot |X| + d(Y, B) \cdot |Y|}{|X| + |Y|}$$

4.1.3 Metoda nejvzdálenějšího souseda

Metoda nejvzdálenějšího souseda vypočítává novou vzdálenost dvou shluků A a B jako maximum všech vzdáleností dvojic prvků a a b takových, že prvek a náleží do shluku A a prvek b náleží do shluku B. [33]

4.2 Nehierarchické shlukovací metody

4.2.1 K-means

Shlukovací metoda K means dostává na vstupu množinu vektorů a počet shluků, který je třeba vytvořit. Nejprve je třeba zvolit počáteční pozice centroidů. Ty mohou být zvoleny náhodně, může být využito apriorních zanlostí o vstupních datech, nebo lze využít algoritmy, které počáteční pozice centroidů odhadnou. Dále probíhá výpočet iterativně. Všechny vektory jsou přiřazeny nejbližšímu centroidu. Následně jsou vypočteny nové centroidy jako geometrické středy shluků. Tyto dva kroky se opakují tak dlouho, dokud po aktualizaci pozic centroidů není žádný vektor, který by změnil svůj shluk. [37, 38, 39]

5 cgMLST analýza

K provedení cgMLST analýzy je třeba nejprve z nasekvenovaných dat sestavit bakteriální genomy. V této práci byla použita sekvenační data z platforem Illumina Miseq a Oxford Nanopore Technologies Minion. Data z Illumina Miseq byla sestavena jak způsobem *de novo* pomocí assembleru SPAdes, verzí 3.14.0 [16], tak zarovnáním k referenci pomocí BWA aligneru, verzí 0.7.3 [17]. Data z ONT Minion byla sestavena *de novo* assemblerem Flye, verzí 2.8 [28]. Dále je třeba zkontrolovat kvalitu sestavených genů.

Po sestavení genomů byly staženy z internetové databáze referenční alely pro cgMLST analýzu. Alely byly vyhledány programem BLAST, verzí 2.2.27, určeny a jejich porovnáním byla vytvořena matice vzdáleností.

Z této matice vzdáleností pak byly vytvořeny grafy umožňující snadnou interpretaci výsledků analýzy.

5.1 Sestavování genomů

Sestavování genomů s využitím dat nasekvenovaných na platformě Illumina bylo provedeno dvěma různými způsoby. Způsobem *de novo* s využitím assembleru SPAdes a skládáním k referenci s využitím aligneru BWA. Nejprve byla čtení vložena do programu FastQC, verze 0.11.9, pomocí kterého byla ohodnocena jejich kvalita. Sledované parametry byly pokrytí referenčního genomu, kvalita mapování k referenčnímu genomu a podíl cytosinu a guaninu. Bylo zjištěno, že čtení jsou vhodná k sestavení. [16]

Při sestavování k referenci byly dále odstřiženy adaptory pomocí programu Trimmomatic [23]. Poté byly pomocí BWA aligneru [24] vytvořeny soubory formátu SAM. Ty byly programem Samtools [20, 22] převedeny na formát BAM. Poté bylo provedeno hodnocení kvality zarovnání pomocí programu Qualimap [25]. Následně byla pomocí Samtools odstraněna čtení, která byla buď nenamapována, nebo nebylo druhé čtení z páru namapováno na odpovídající pozici. Následně byl soubor formátu BAM převeden pomocí BCFtools na formát VCF a z toho byla pomocí vcfutils [21] vytvořena konsensuální sekvence ve formátu FASTQ. Z té byl pomocí programu seqtk vytvořen soubor FASTA a to tak, že báze s PHRED skóre větším než 20 byly ponechány, ostatní byly označeny za neurčené nahrazením N.

Sestavování *de novo* bylo provedeno s využitím programu SPAdes[16], spuštěného s parametrem –isolate. Program SPAdes provedl jak korekci chybných čtení pomocí Bayeshammer, tak následné sestavení genomů. Kvalita výsledného sestavení byla hodnocena programem Quast. [26]

Sestavování dat z přístroje MinION probíhalo s využitím programů Guppy[18] a Flye [28]. Program Guppy byl využit k basecallingu a demultiplexaci. K následnému složení genomů byl využit program Flye. Hodnocení kvality sestavených genomů proběhlo s využitím programu Quast. [26]

5.2 Hodnocení kvality

5.2.1 Hodnocení čtení z Illumina Miseq pomocí FastQC

Kvalita všech čtení ze sekvenátoru Illumina Miseq byla ohodnocena programem FastQC[27], konkrétně jeho webovou verzí. Byly sledovány zejména parametry kvalita sekvence podle bazí, skóre kvality sekvence podle bazí a přítomnost adaptorů. Byla potvrzena vhodnost všech vzorků pro sestavení a přítomnost adaptorů v některých vzorcích. Z přítomnosti adaptorů plyne potřeba jejich odstranění.

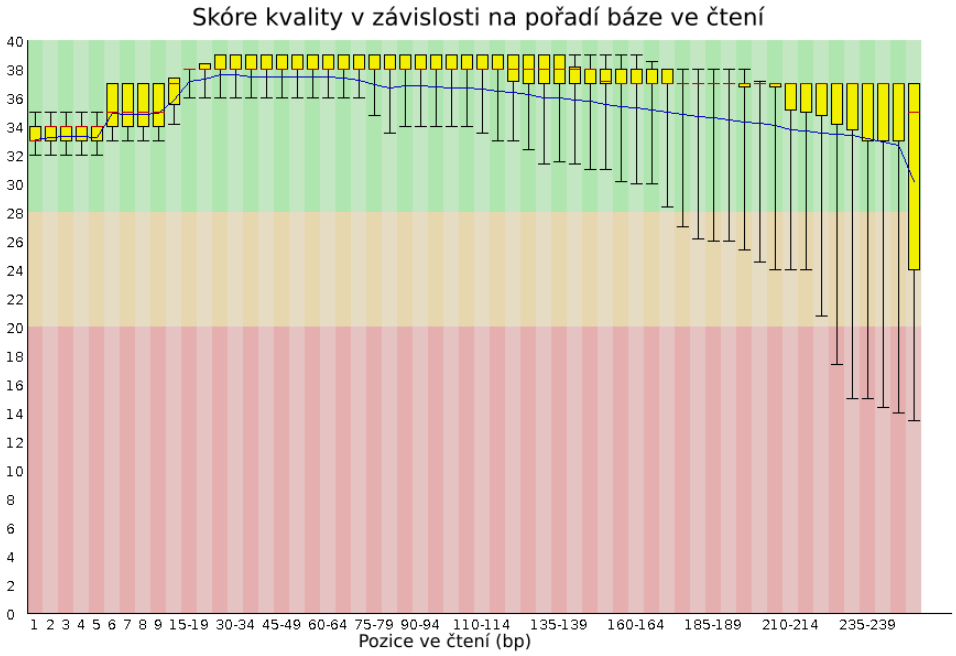
Na obrázku 5.1 je zobrazen graf rozložení PHRED skóre na dané pozici ve všech čteních. Na ose x je pořadí báze ve čtení, na ose y pak PHRED skóre. Červená čára označuje mediánové PHRED skóre na dané pozici, modrá čára reprezentuje průměrné PHRED skóre na dané pozici. Žlutý box ohraničuje PHRED skóre dvou vnitřních kvartilů, černé linky pak oddělují horní a dolní decil.

Obrázek 5.2 představuje graf rozložení podílu cytosinu a guaninu ve čtení. Na ose x je podíl cytosinu a guaninu, tedy součet výskytu bazí cytosin a guanin, vydělený celkovou délkou čtení, vyjádřený v procentech. Na ose y je pak počet čtení s odpovídajícím podílem cytosinu a guaninu.

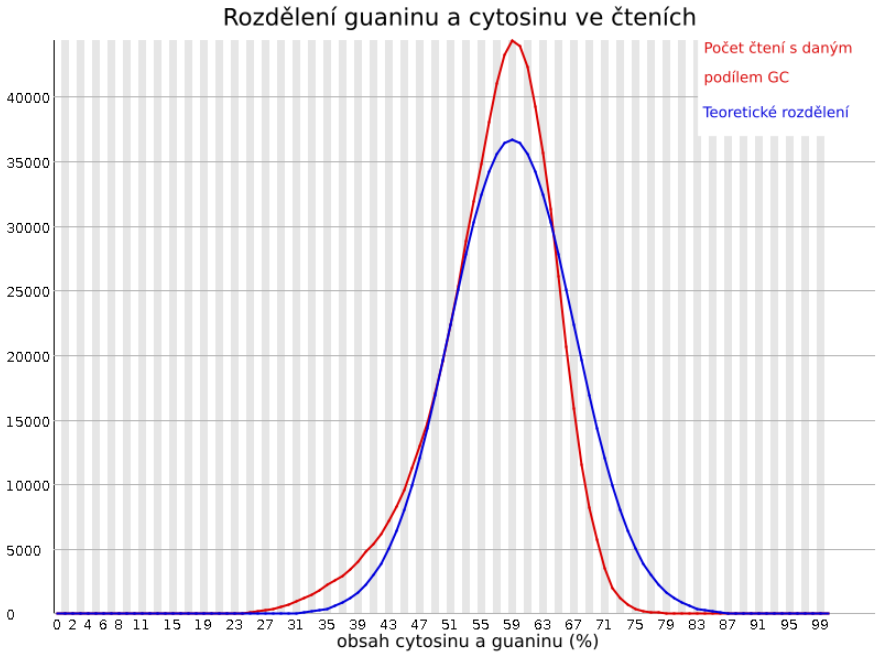
Na obrázku 5.3 je graf znázorňující přítomnost adaptorů.

5.2.2 Hodnocení kvality zarovnání k referenci programem Qualimap

Program Qualimap [25]byl využit k hodnocení kvality zarovnání k referenci. Na obrázku 5.4 lze vidět graf pokrytí referenčního genomu čteními jednotlivých vzorků. Na ose x je pozice v referenčním genomu, na ose y pokrytí čteními daného vzorku.

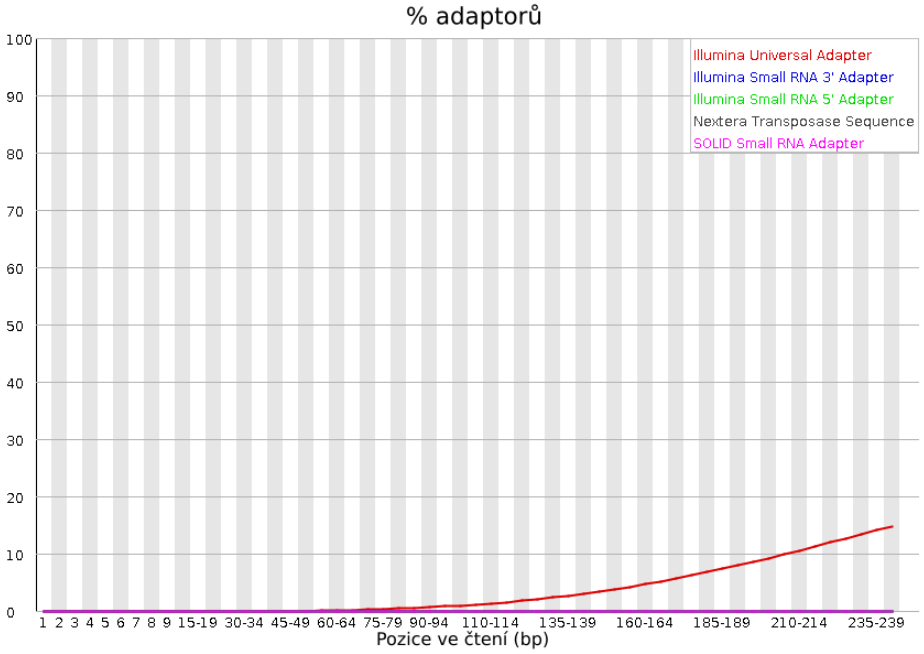


Obr. 5.1: Graf PHRED skóre v závislosti na pořadí báze ve čtení

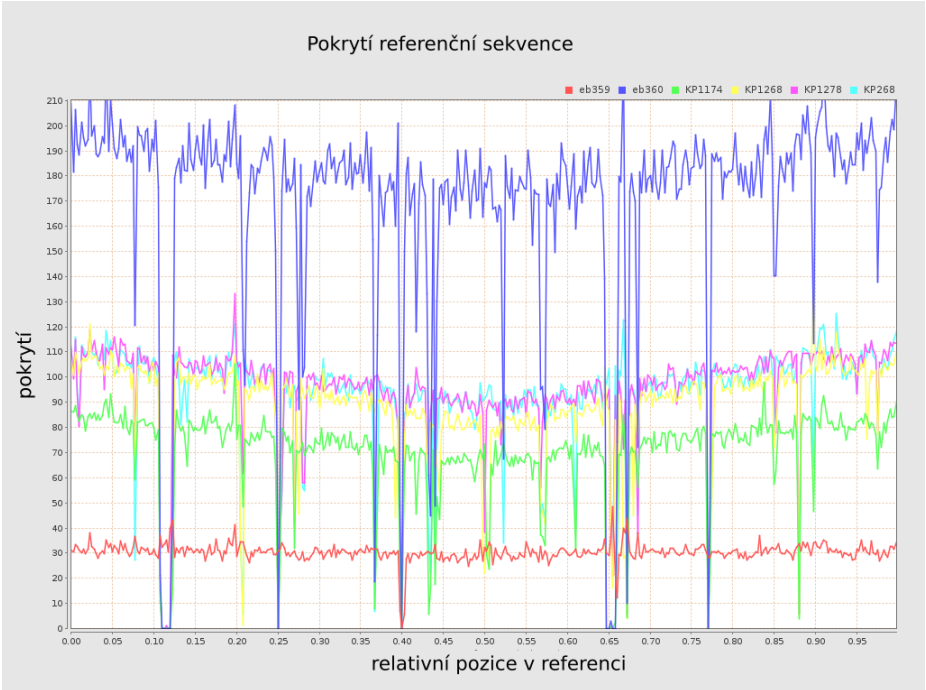


Obr. 5.2: Graf znázorňující procento guaninu a cytosinu ve čteních

Dalšími hodnocenými parametry byly průměrné pokrytí a procento guaninu a cytosinu. Všechny genomy složené k referenci byly vyhodnoceny jako vhodné pro další analýzu.



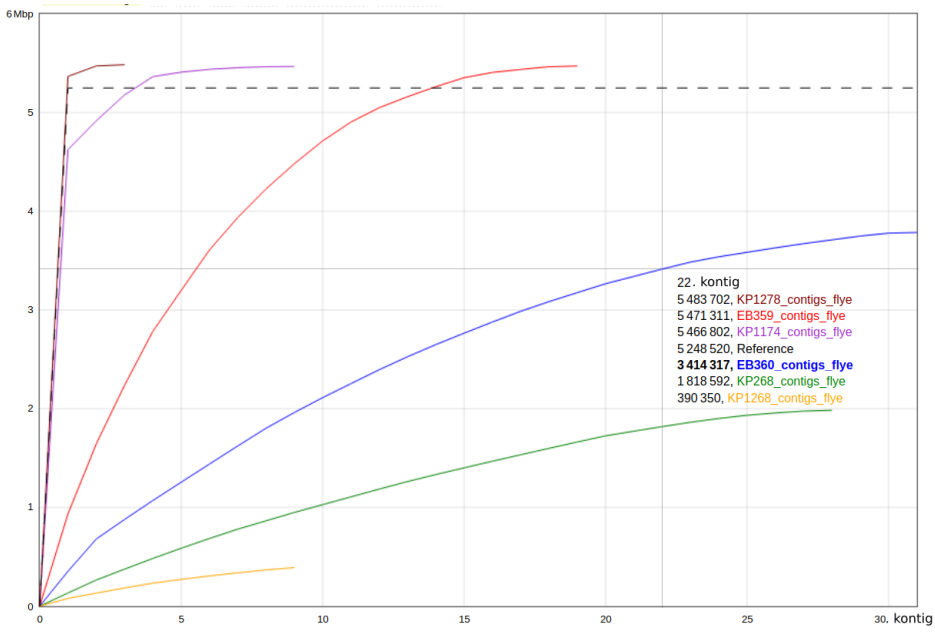
Obr. 5.3: Graf zázorňující přítomnost adaptorů ve čteních.



Obr. 5.4: Graf pokrytí referenčního genomu

5.2.3 Hodnocení kvality skládání *de novo* programem Quast

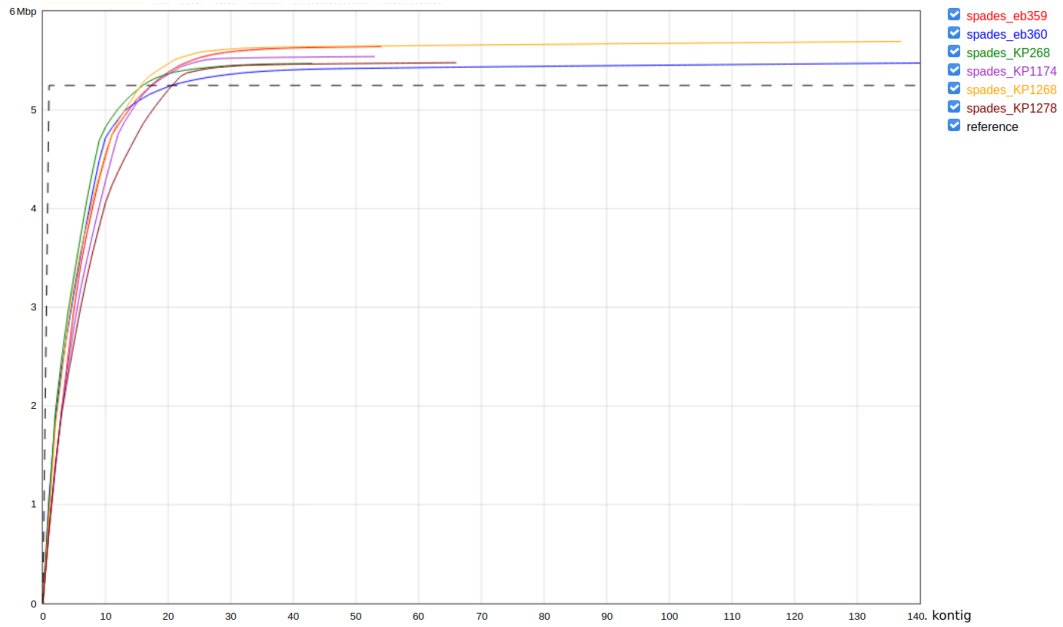
Program Quast [26] byl použit k hodnocení kvality skládání *de novo* a to jak u genomů poskládaných ze čtení z Illumina Miseq, tak u genomů poskládaných ze čtení z Oxford Nanopore Technologies MinION. Hodnocenými parametry byly například N50, L50 a počet kontigů. V grafu na obrázku 5.5 lze vidět součet délky nejdelších x kontigů v závislosti na x pro jednotlivé genomy sestavené programem Flye [28] ze čtení z Oxford Nanopore Technologies MinION. V grafu na obrázku 5.6 pak lze vidět stejný parametr pro genomy sestavené pomocí programu SPAdes [16] z dat ze sekvenátoru Illumina Miseq. Na základě tohoto hodnocení kvality, zejména grafu 5.5 byly vyřazeny z další analýzy genomy KP1268 a KP268 sestavené ze čtení sekvenátoru Oxford Nanopore Technologies MinION. Všechny genomy sestavené pomocí programu SPAdes s využitím čtení ze sekvenátoru Illumina Miseq byly vyhodnoceny jako vhodné pro další analýzu.



Obr. 5.5: Součet délek nejdelších x kontigů pro sesavení s využitím dat z Oxford Nanopore Technologies MinION

5.3 Vytvoření matice vzdáleností

Část analýzy, ve které jsou v již sestavených genomech nalezeny referenční alely, vybrány vhodné geny a vytvořena matice vzdáleností je znázorněna blokovým schématem na obrázku 5.7.



Obr. 5.6: Součet délek nejdelších x kontigů pro sestavení s využitím dat z Illumina Miseq

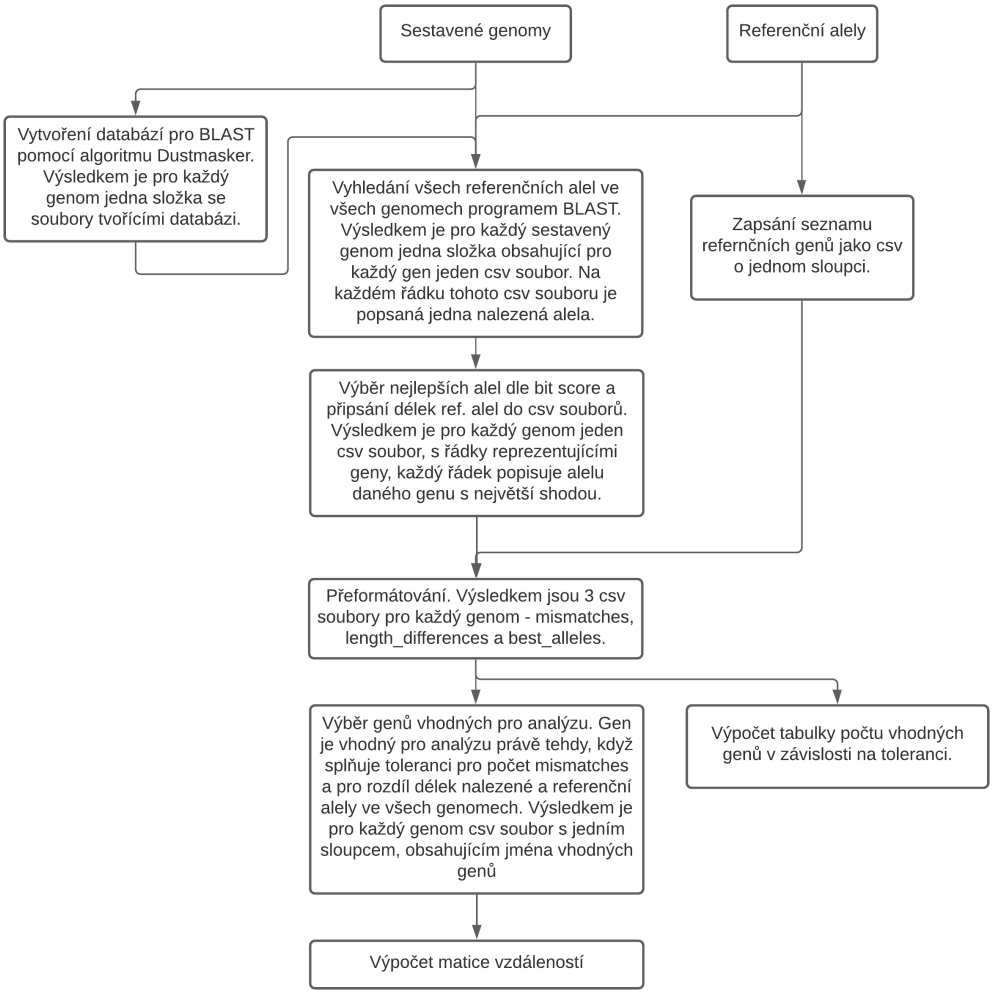
Vstupy algoritmu jsou sestavené genomy, jejichž získání bylo popsáno výše a referenční alely stažené z internetové databáze. [30, 31, 32]

5.3.1 Vyhledání referenčních genů v genomech

Nejprve bylo třeba vytvořit databáze pro program BLAST. Následně byl program BLAST spuštěn tak, že vyhledával všechny referenční alely všech referenčních genů ve všech sestavených genomech. Poté byla pro každou dvojici sestaveného genomu a referenčního genu vybrána a ponechána nalezená alela s nejvyšším bit score, ostatní nalezené alely nebyly dále využity. Byly vypočteny délky nalezených referenčních alel.

5.3.2 Výběr alel vhodné kvality

Následně bylo třeba vybrat alely, které se dostatečně shodovaly s referenčními. Byla využita dvě kritéria, počet neshodných nukleotidů a rozdíl délek nalezené alely a referenční alely. Aby byla kritéria správně nastavena, bylo pro několik dvojic maximálního počtu neshodných nukleotidů a maximálního rozdílu délek referenční a nalezené alely vypočteno, kolik alel by danou dvojici kritérií splňovalo v každém genomu, a kolik genů by po vyřazení alel nesplňujících kritéria bylo nalezeno ve



Obr. 5.7: Blokové schéma znázorňující postup pro porovnání genomů

všech genomech sestavených daným assemblerem. Výsledky byly zaneseny do tabulek, jedné pro každý assembler.

Bylo zjištěno, že pro genomy sestavené assemblery BWA a SPAdes není třeba povolovat žádné neshodné nukleotidy ani žádný rozdíl délek nalezené a referenční alely. I po vyřazení všech alel, které se přesně neshodovaly s referenčními zbyl dostatek genů pro provedení kvalitní analýzy. V analýze genomů sestavených assemblerem SPAdes, bylo pro analýzu ponecháno 2333 genů z původních 2358, u genomů sestavených assemblerem BWA bylo ponecháno 987 genů.

Pro genomy sestavené assemblerem Flye byla situace odlišná. Pokud by byly ponechány jen přesně nalezené alely a vyřadili všechny geny, kde u některého z genomů

nebyla alela nalezena v dostatečné kvalitě, stejně jako u ostatních assemblerů, zbylo by pro cgMLST analýzu z původních 2358 genomů jen 72. To by bylo pro provedení kvalitní analýzy zcela nedostatečné. Dále bylo zjištěno, že v genomu vzorku EB360 sestaveném pomocí Flye, bylo výrazně méně kvalitních alel, než v ostatních genomech sestavených tímto assemblerem. Toto zjištění koresponduje s výsledky hodnocení sestavených genomů.

V následujících tabulkaách značí podmínka 1 požadavek, aby v nalezené alele nebyly žádné mismatche a délka se přesně shodovala s délkou referenční alely. Podmínka 2 značí požadavek, aby v nalezené alele byl maximálně jeden mismatch a zachování požadavku na shodnost délek alel. Podmínka 3 povoluje až tři mismatche a rozdíl délek nalezené a referenční alely maximálně jedna. Podmínka 4 pak dooluje až pět mismatchů a rozdíl délek nalezené a referenční alely také až pět.

Tab. 5.1: Tabulka počtu kvalitních alel v závislosti na toleranci pro genomy sestavené assemblerem Flye

	Podmínka 1	Podmínka 2	Podmínka 3	Podmínka 4
EB359	546	772	1309	1746
EB360	285	437	763	1133
KP1174	1124	1512	2038	2268
KP1278	1289	1604	2090	2286
všechny	72	134	409	823

Tab. 5.2: Tabulka počtu kvalitních alel v závislosti na toleranci pro genomy sestavené assemblerem SPAdes

	Podmínka 1	Podmínka 2	Podmínka 3	Podmínka 4
EB359	2352	2352	2352	2352
EB360	2350	2350	2353	2353
KP1174	2355	2355	2356	2356
KP1268	2352	2352	2355	2356
KP1278	2356	2356	2357	2357
KP268	2355	2355	2355	2356
všechny	2333	2333	2338	2340

Tab. 5.3: Tabulka počtu kvalitních alel v závislosti na toleranci pro genomy sestavené assemblerem BWA

	Podmínka 1	Podmínka 2	Podmínka 3	Podmínka 4
EB359	1024	1580	2071	2248
EB360	2317	2321	2329	2340
KP1174	2310	2324	2331	2337
KP1268	2296	2310	2322	2331
KP1278	2326	2335	2336	2343
KP268	2283	2317	2325	2335
všechny	987	1530	2016	2204

Kvůli nekvalitně sestavenému vzorku EB360 byla analýza genomů sestavených s pomocí assembleru Flye dále rozdělena na dvě. V jedné jsou dále analyzovány jen genomy EB359, KP1174 a KP1278, ve druhé je spolu s nimi ponechán i genom EB360. Samotné vyřazení nejméně kvalitního genomu však nestačilo k získání dostatečného množství alel, v obou analýzách bylo třeba nastavit kritéria kvalitní alely benevolentněji, než u předchozích assemblerů. Pro analýzu se třemi genomy byla použita podmínka 3, byly tedy povoleny maximálně tři rozdílné nukleotidy a rozdíl délek referenční a nalezené alely maximálně jedna. Pro analýzu se čtyřmi genomy byla použita podmínka 4, tedy bylo povoleno až pět rozdílných nukleotidů a rozdíl délek referenční a nalezené alely také maximálně pět. Takto zbylo v analýze tří genomů 1060 alel a v analýze čtyř genomů 823 alel.

5.3.3 Výpočet matice vzdáleností

Následoval výpočet matice vzdáleností. Vzdálenost dvou genomů byla vypočtena porovnáváním nalezených referenčních alel. Za každý gen u kterého byly ve dvou genomech nalezeny různé alely byla vzdálenost těchto dvou genomů zvětšena o jedna. Výsledkem jsou celá čísla v rozmezí 0 až počet genů použitých pro analýzu daných genomů.

5.4 Grafické zobrazení výsledků

Z tabulek vzdáleností byly vykresleny minimální kostry grafu (MST z angl. minimum spanning tree) a dendrogramy metodami UPGMA a shlukováním nejvzdálenějšího souseda. Jak již bylo zmíněno v kapitole 5.3.2, pro assembler Flye byly provedeny dvě analýzy, jedna se třemi vzorky, využívající více referenčních genů, druhá se čtyřmi vzorky.

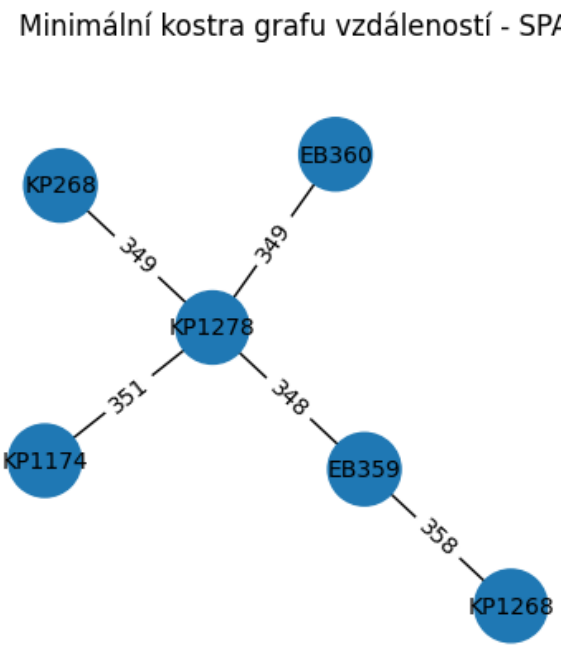
6 Zhodnocení výsledků

Minimální kostry grafů vzdáleností genomů sekvenovaných na platformě Illumina, sestavovaných assemblery SPAdes, v grafu 6.1 a BWA, v grafu A.1, se liší jen vzdálenostmi mezi genomy, nikoliv uspořádáním genomů v grafu. Výsledek odpovídá tomu, že genomy sestavené assemblerem SPAdes, se dobře shodují s genomy sestavenými assemblerem BWA. Genomy nasekvenované na platformě Oxford Nanopore Technologies MinION (ONT), nebyly všechny sestaveny v dostatečné kvalitě a proto na nich byly provedeny dvě analýzy, jedna se třemi a druhá se čtyřmi genomy. U MST genomů nasekvenovaných na platformě ONT, v grafech A.2 a A.3, nemůžeme určit jestli si vzájemně odpovídají, protože v analýze na čtyřech genomech má ke všem ostatním genomům nejblíže genom EB360, který se u analýzy na třech genomech nevyskytuje. Při porovnání MST genomů sestavených z dat sekvenovaných na platformě Illumina a MST genomů sestavovaných z dat z ONT je patrné, že si grafy navzájem neodpovídají.

U analýz UPGMA na grafech 6.2, A.4, A.5, A.6 a shlukováním meotdou nejvzálenějšího souseda na grafech 6.3, A.7, A.8, A.9je patrné, že si dendrogramy neodpovídají ani v rámci dat z jedné sekvenační platformy, natož pak napříč platformami. Příčina, proč si analýzy neodpovídají je nejlépe patrná v MST genomů sestavených assemblerem SPAdes na grafu 6.1. Hned tři nejmenší vzdálenosti genomů (EB359-KP1278, EB360-KP1278, KP268-KP1278) se vzájemně liší maximálně o jednu alelu. Čtvrtá nejmenší vzdálenost (KP1174-KP1278) se od nich liší maximálně o tři alely. Vzdálenosti jsou si tedy velmi podobné. Dendrogramům vytvořeným hierarchickými shlukovacími metodami je proto třeba nepřikládat příliš velkou váhu. Pokud bychom nalezli jen o několik genomů více, nebo méně, nebo bychom určili některé alely špatně, je velmi pravděpodobné, že by výsledné dendrogramy byly jiné.

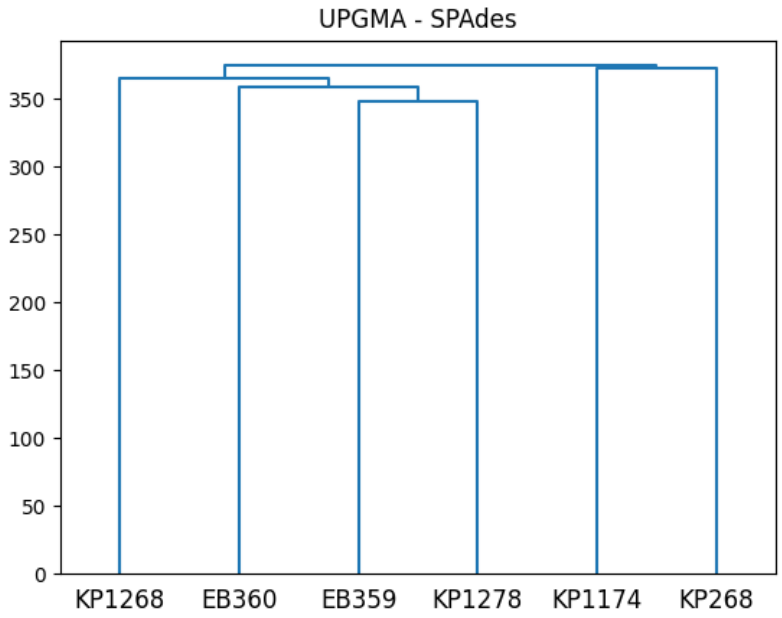
Na MST genomů sestavených assemblerem BWA, zobrazeném na grafu A.1, jsou rozdíly mezi vzdálenostmi genomů o něco vyšší. Stále jsou však poměrně malé a dendrogramům vytvořeným z těchto genomů je také třeba nepřikládat příliš velkou váhu. V MST analýzy čtyř genomů sestavených pomocí assembleru Flye, je patrné, že nejmenší vzdálenost mezi genomy není unikátní. Vzdálenost 142 mají jak KP1278 s EB360, tak KP1174 s EB360. Je tedy zřejmé, že již první krok metody UPGMA i metody nejvzdálenějšího souseda není jednoznačný.

V dendrogramech vytvořených hierarchickými shlukovacími metodami tedy nejsou patrné výrazné podobnosti a z MST vykazují výraznou vzájemnou podobnost jen

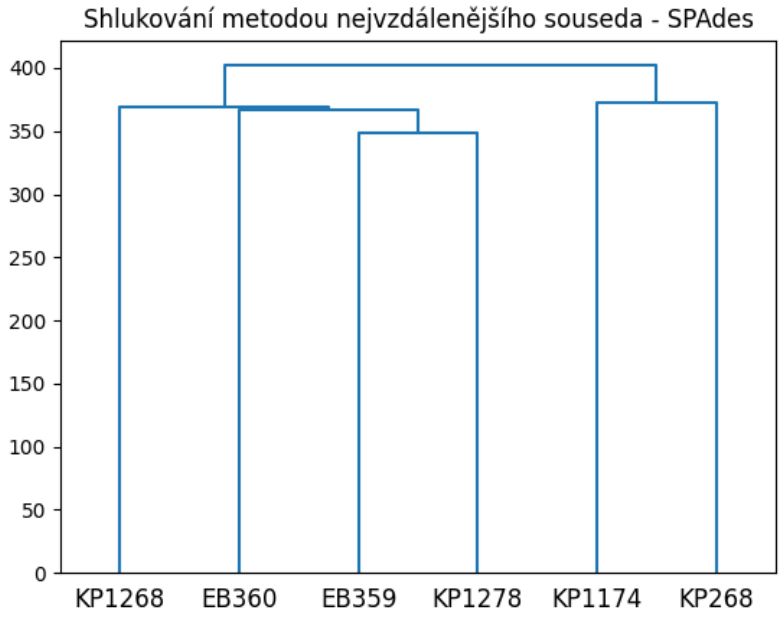


Obr. 6.1: Minimální kostra grafu vzdáleností genomů sestavených assemblerem SPAdes

MST genomů sestavených assemblerem SPAdes a MST z genomů sestavených assemblerem BWA. Všechny metody však dospěly ke společnému závěru, že vzorky si nejsou vzájemně geneticky podobné a nelze v nich pozorovat jednoznačné shluky.



Obr. 6.2: Dendrogram vytvořený metodou UPGMA z genomů sesavených assemblerem SPAdes



Obr. 6.3: Dendrogram vytvořený metodou nejvzálenějšího souseda z genomů sestavených assemblerem SPAdes

Závěr

V teoretické části práce byla popsána sekvenace genomu a vybrané sekvenační přístroje. Ze druhé generace sekvenátorů jsou v práci obsaženy platformy Roche 454 a Illumina. Ze třetí generace sekvenátorů jsou pak popsány přístroje firem Pacific biosciences a Oxford Nanopore Technologies. Dále bylo v teoretické části práce popsáno sestavování genomů, Debruijnovy grafy a metoda OLC.

Z obdržených sekvenačních dat šesti různých bakterií Klepsiella pneumoniae sekvenovaných na dovu platformách, konkrétně Illumina Miseq a Oxford Nanopore Technolgies MinION byly sestaveny genomy. S využitím dat z přístroje Illumina Miseq byly genomy sestaveny dvěma různými způsoby a to *de novo* a skládáním k referenci. Čtení ze sekvenátoru MinION byly využity pouze k sestavení *de novo*. Dále byla zhodnocena kvalita sestavení s využitím programů FastQC[27], Quast[26] a Qualimap[25]. Dva genomy musely být na základě tohoto hodnocení kvality vyřazeny z další analýzy.

Dále byl navržen a implementován algoritmus pro vyhledávání genů, který využívá programu BLAST. Byly vybrány geny nalezené v dostatečné kvlaitě a provedena cgMLST analýza. Výsledky byly graficky zobrazeny a to s využitím minimálních koster grafů vzdáleností a shlukových analýz vytvořených metodou UPGMA a metodou nejvzálenějšího souseda.

Vzájemně si odpovídaly vykreslení MST u genomů sestavených assemblery SPAdes a BWA. Vzdálenosti genomů byly mezi sebou velmi podobné a proto je výsledkem analýzy, že vzorky nelze jednoznačně rozdělit do dobře vypovídajících shluků.

Literatura

[1] RUPPITSCH, W. Molecular typing of bacteria for epidemiological surveillance and outbreak investigation. Bodenkultur [online]. De Gruyter Open, 2016, 67(4), 199-224 [cit. 2020-11-27]. ISSN 00065471. Dostupné z: doi:10.1515/boku-2016-0017

[2] EL-METWALLY, Sara, Osama M. OUDA a Mohamed HELMY. *Next generation sequencing technologies and challenges in sequence assembly.* New York: Springer, c2014. SpringerBriefs in system biology. ISBN 978-1-4939-0714-4.

[3] VOELKERDING, Karl V, Shale A DAMES, Jacob D DURTSCHI a Karl V VOELKERDING. Next-generation sequencing: from basic research to diagnostics. Clinical chemistry [online]. 2009, 55(4), 641-658 [cit. 2020-11-30]. ISSN 00099147. Dostupné z: doi:10.1373/clinchem.2008.112789

[4] ONMUS-LEONE, Fatma, Jun HANG, Robert J CLIFFORD, Yu YANG, Matthew C RILEY, Robert A KUSCHNER, Paige E WATERMAN a Emil P LESHO. E*nhanced de novo assembly of high throughput pyrosequencing data using whole genome mapping.* PloS one [online]. 2013, 8(4), e61762 [cit. 2020-11-28]. ISSN 1932-6203. Dostupné z: doi:10.1371/journal.pone.0061762

[5] DAMES, Shale. Pyrosequencing. Researchgate [online]. [cit. 2021-01-28]. Dostupné z: https://www.researchgate.net/figure/Roche-454-GS-FLX-sequencing-Template-DNA-is-fragmented-end-repaired-ligated-to_fig1_24043867

[6] HODKINSON, Brendan P, Elizabeth A GRICE a Brendan P HODKINSON. Next-Generation Sequencing: A Review of Technologies and Tools for Wound Microbiome Research. Advances in wound care [online]. 2015, 4(1), 50-58 [cit. 2020-11-18]. ISSN 2162-1918. Dostupné z: doi:10.1089/wound.2014.0542

[7] Illumina sequencing platforms. Illumina [online]. [cit. 2021-01-01]. Dostupné z: https://www.illumina.com/systems/sequencing-platforms.html

[8] LU, Yuan, Yingjia SHEN, Wesley WARREN a Ronald WALTER. Next Generation Sequencing in Aquatic Models. Next Generation Sequencing - Advances, Applications and Challenges. InTech, 2016, 2016-01-14. ISBN 978-953-51-2240-1. Dostupné z: doi:10.5772/61657

[9] AU, Kin Fai, Jason G UNDERWOOD, Lawrence LEE, Wing Hung WONG a Yi XING. Improving PacBio Long Read Accuracy by Short Read Alignment (Error Correction of PacBio Long Read) [online]. San Francisco, USA:

Public Library of Science, 2012, 7(10), e46679 [cit. 2020-11-30]. Dostupné z: doi:10.1371/journal.pone.0046679

[10] RHOADS, Anthony a Kin Fai AU. PacBio Sequencing and Its Applications. Genomics, proteomics & bioinformatics [online]. Elsevier, 2015, 13(5), 278-289 [cit. 2020-11-30]. ISSN 1672-0229. Dostupné z: doi:10.1016/j.gpb.2015.08.002

[11] LU, Hengyun, Francesca GIORDANO a Zemin NING. *Oxford Nanopore MinION Sequencing and Genome Assembly. Genomics, proteomics & bioinformatics* [online]. Elsevier, 2016, 14(5), 265-279 [cit. 2020-11-18]. ISSN 1672-0229. Dostupné z: doi:10.1016/j.gpb.2016.05.004

[12] Oxford Nanopore Technologies, Product brochure [online] [cit. 2020-11-29] Dostupné z: https://nanoporetech.com/sites/default/files/s3/literature/product-brochure.pdf

[13] BRYANT, Rd a H FREDRICKSEN. COVERING THE DEBRUIJN GRAPH. Discrete Mathematics [online]. ELSEVIER SCIENCE BV, 1991, 89(2), 133-148 [cit. 2020-11-18]. ISSN 0012-365X. Dostupné z: doi:10.1016/0012-365X(91)90362-6

[14] LANGMEAD, Ben. De Bruijn Graph assembly [online]. In: . [cit. 2021-01-05]. Dostupné z: https://www.cs.jhu.edu/ langmea/resources/lecture_notes/assembly_dbg.pdf

[15] LANGMEAD, Ben. Overlap Layout Consensus assembly [online]. In: . [cit. 2021-01-21]. Dostupné z: http://www.cs.jhu.edu/ langmea/resources/lecture_notes/assembly_olc.pdf

[16] NURK, S., A. BANKEVICH, D. ANTIPOV, et al. Assembling genomes and mini-metagenomes from highly chimeric reads. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) [online]. 2013, s. 158-170 [cit. 2021-01-05]. ISBN 9783642371943. ISSN 03029743. Dostupné z: doi:10.1007/978-3-642-37195-0_13

[17] Manual Reference Pages - bwa (1) [online]. 8.3.2013 [cit. 2021-01-05]. Dostupné z: http://bio-bwa.sourceforge.net/bwa.shtml

[18] Guppy software overview [online]. [cit. 2021-01-06]. Dostupné z: https://community.nanoporetech.com/protocols/Guppy-protocol/v/gpb_2003_v1_revu_14dec2018/guppy-software-overview

[19] SCHÜRCH, A.C, S ARREDONDO-ALONSO, R.J.L WILLEMS a R.V GOERING. Whole genome sequencing options for bacterial strain typing and epidemiologic analysis based on single nucleotide polymorphism versus gene-by-gene–based approaches. Clinical microbiology and infection [online]. Elsevier, 2018, 24(4), 350-354 [cit. 2021-5-25]. ISSN 1198-743X. Dostupné z: doi:10.1016/j.cmi.2017.12.016

[20] LI H, HANDSAKER B, WYSOKER A, FENNELL T, RUAN J, HOMER N, MARTH G, ABECASIS G, DURBIN R, and 1000 Genome Project Data Processing Subgroup, The Sequence alignment/map (SAM) format and SAMtools, Bioinformatics (2009) 25(16) 2078-9 [19505943]

[21] DANECEK P, AUTON A, ABECASIS G, ALBERS CA, BANKS E, DEPRISTO MA, HANDSAKER RE, LUNTER G, MARTH GT, SHERRY ST, MCVEAN G, DURBIN R, 1000 Genomes Project Analysis Group, The variant call format and VCFtools, Bioinformatics (2011) 27(15) 2156-8 [21653522]

[22] LI H, A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data, Bioinformatics (2011) 27(21) 2987-93. [21903627]

[23] BOLGER, A. M., LOHSE, M., & USADEL, B. (2014). Trimmomatic: A flexible trimmer for Illumina Sequence Data. Bioinformatics, btu170.

[24] LI H. and DURBIN R. (2010) Fast and accurate long-read alignment with Burrows-Wheeler Transform. Bioinformatics, Epub. [PMID: 20080505]

[25] Konstantin Okonechnikov, Ana Conesa and Fernando García-Alcalde "Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data."Bioinformatics(2015)

[26] GUREVICH, Alexey, Vladislav SAVELIEV, Nikolay VYAHHI a Glenn TESLER. QUAST: quality assessment tool for genome assemblies. Bioinformatics [online]. Oxford University Press, 2013, 29(8), 1072-1075 [cit. 2021-01-05]. ISSN 1367-4803. Dostupné z: doi:10.1093/bioinformatics/btt086

[27] ANDREWS, S. (2010). FastQC: A Quality Control Tool for High Throughput Sequence Data [online]. Dostupné z: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

[28] Flye [online]. [cit. 2021-01-31]. Dostupné z: https://github.com/fenderglass/Flye

[29] LI, Wenjun, Didier RAOULT a Pierre-edouard FOURNIER. Bacterial strain typing in the genomic era. FEMS Microbiology Reviews [online]. Blackwell Publishing, 2009, 33(5), 892-916 [cit. 2021-5-24]. ISSN 01686445. Dostupné z: doi:10.1111/j.1574-6976.2009.00182.x

[30] cgMLST.org Nomenclature Server. [online].[cit. 2021-04-31]. Dostupné z: https://www.cgmlst.org/ncs

[31] WEBER, Robert E, Michael PIETSCH, Andre FRÜHAUF, et al. IS-Mediated Transfer of as the Main Route of Resistance Transmission During a Polyclonal, Multispecies Outbreak in a German Hospital. Frontiers in microbiology [online]. 2019, 10, 2817 [cit. 2021-5-21]. ISSN 1664-302X. Dostupné z: doi:10.3389/fmicb.2019.02817

[32] PIAZZA, Aurora, Francesco COMANDATORE, Francesca ROMERI, et al. Identification of Gene in ST307 and ST661 Clones in Italy: Old Acquaintances for New Combinations. Microbial drug resistance (Larchmont, N.Y.) [online]. 2019, 25(5), 787 [cit. 2021-5-21]. ISSN 1076-6294. Dostupné z: doi:10.1089/mdr.2018.0327

[33] MURTAGH, Fionn a Pedro CONTRERAS. Algorithms for hierarchical clustering: an overview, II. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery [online]. Hoboken, USA: Wiley Periodicals, 2017, 7(6), n/a-n/a [cit. 2021-5-22]. ISSN 1942-4787. Dostupné z: doi:10.1002/widm.1219

[34] MÜLLNER, D. Fastcluster: Fast hierarchical, agglomerative clustering routines for R and Python. Journal of Statistical Software [online]. American Statistical Association, 2013, 53(9), 1-18 [cit. 2021-5-23]. ISSN 15487660. Dostupné z: doi:10.18637/jss.v053.i09

[35] KAPETANOVIC, Izet M., Simon ROSENFELD a Grant IZMIRLIAN. Overview of Commonly Used Bioinformatics Methods and Their Applications. Annals of the New York Academy of Sciences [online]. Oxford, UK: Blackwell Publishing, 2004, 1020(1), 10-21 [cit. 2021-5-23]. ISSN 0077-8923. Dostupné z: doi:10.1196/annals.1310.003

[36] DAVIDSON, Ruth a Seth SULLIVANT. Polyhedral combinatorics of UPGMA cones. Advances in Applied Mathematics [online]. Elsevier B.V, 2013, 50(2), 327 [cit. 2021-03-19]. ISSN 0196-8858.

[37] KOSLICKI, David, Saikat CHATTERJEE, Damon SHAHRIVAR, Mikko VEHKAPERÄ, Yueheng LAN a Jukka CORANDER. ARK: Aggregation of Reads

by K-Means for Estimation of Bacterial Community Composition. PloS one [online]. Public Library of Science (PLoS), 2015, 10(10), e0140644 [cit. 2021-5-23]. Dostupné z: doi:10.1371/journal.pone.0140644

[38] OSAMOR, Victor Chukwudi, Ezekiel Femi ADEBIYI, Jelilli Olarenwaju OYELADE, Seydou DOUMBIA a Jérémie BOURDON. Reducing the Time Requirement of k-Means Algorithm (Reducing the Time Requirement of k-Means Algorithm) [online]. San Francisco, USA: Public Library of Science, 2012, 7(12), e49946 [cit. 2021-5-23]. Dostupné z: doi:10.1371/journal.pone.0049946

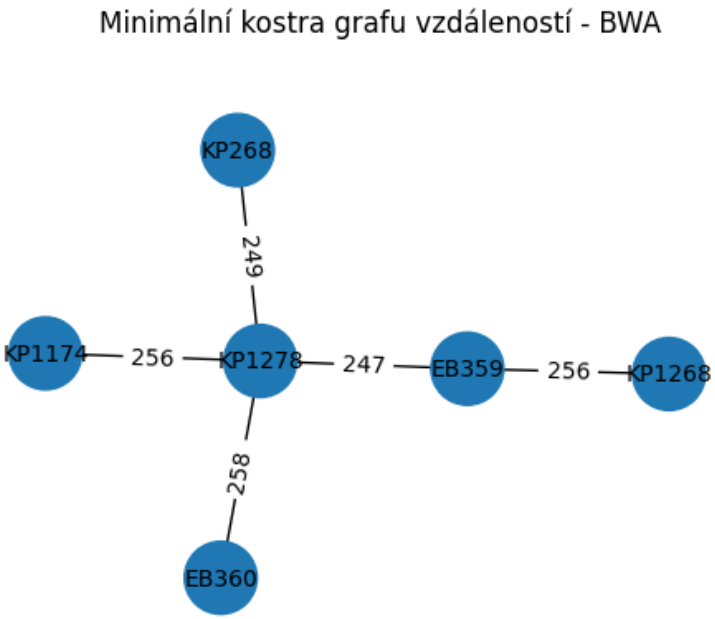
[39] VIRTANEN, PAULI, RALF GOMMERS, TRAVIS E. OLIPHANT, MATT HABERLAND, TYLER REDDY, DAVID COURNAPEAU, EVGENI BUROVSKI, PEARU PETERSON, WARREN WECKESSER, JONATHAN BRIGHT, STÉFAN J. VAN DER WALT, MATTHEW BRETT, JOSHUA WILSON, K. JARROD MILLMAN, NIKOLAY MAYOROV, ANDREW R. J. NELSON, ERIC JONES, ROBERT KERN, ERIC LARSON, C J CAREY, İLHAN POLAT, YU FENG, ERIC W. MOORE, JAKE VANDERPLAS, DENIS LAXALDE, JOSEF PERKTOLD, ROBERT CIMRMAN, IAN HENRIKSEN, E. A. QUINTERO, CHARLES R. HARRIS, ANNE M. ARCHIBALD, ANTÔNIO H. RIBEIRO, FABIAN PEDREGOSA and PAUL VAN MULBREGT, 2020. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nature Methods [online]. vol. 17, no. 3, pp. 261-272 doi:10.1038/s41592-019-0686-2

[40] John D. HUNTER. Matplotlib: A 2D Graphics Environment, Computing in Science & Engineering, 9, 90-95 (2007), DOI:10.1109/MCSE.2007.55

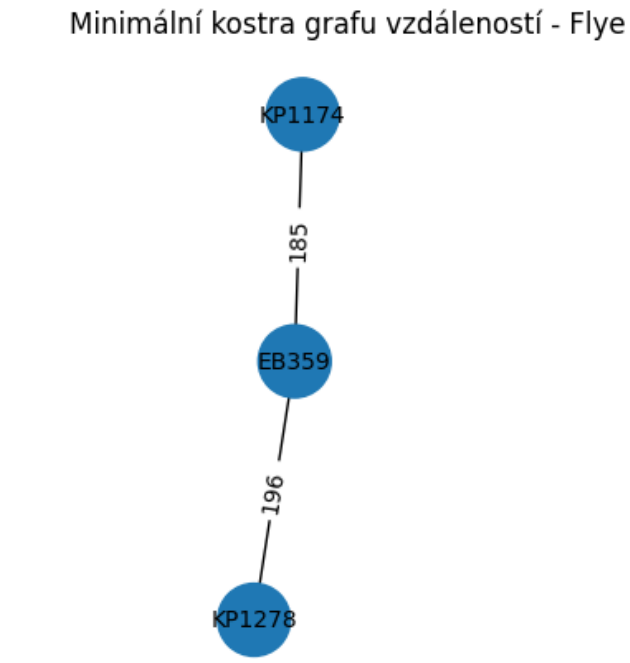
Seznam symbolů, veličin a zkratek

bp pár bazí
BWA aligner Burrows-Wheeler
CCD zařízení s vázanými náboji
DNA deoxyribonukleonová kyselina
dNTP deoxynukleotrifosfát
cgMLST core genome multilokusová sekvenční typizace
MLST multilokusová sekvenční typizace
OLC metoda překrytí, rozložení a shody (angl. Overlap layout consensus)
PCR polymerázová řetězová reakce
SMRT jednomolekulární v reálném čase
SNP jednonukleotidový polymorfismus
SNV jednonukleotidová variace
UPGMA metoda neváženého párování s aritmetickým průměrem
WGS celogenomové sekvenování

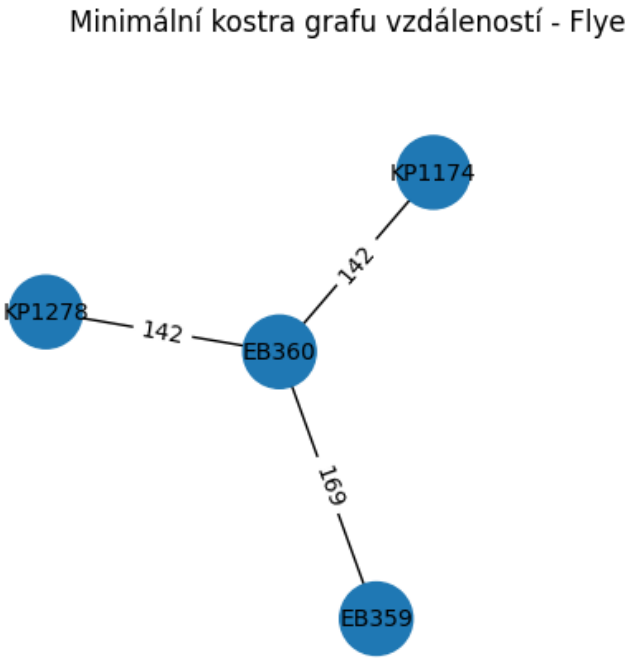
A Grafické vykreslení výsledků ccgMLST analýzy pro genomy sestavené assemblery BWA a Flye



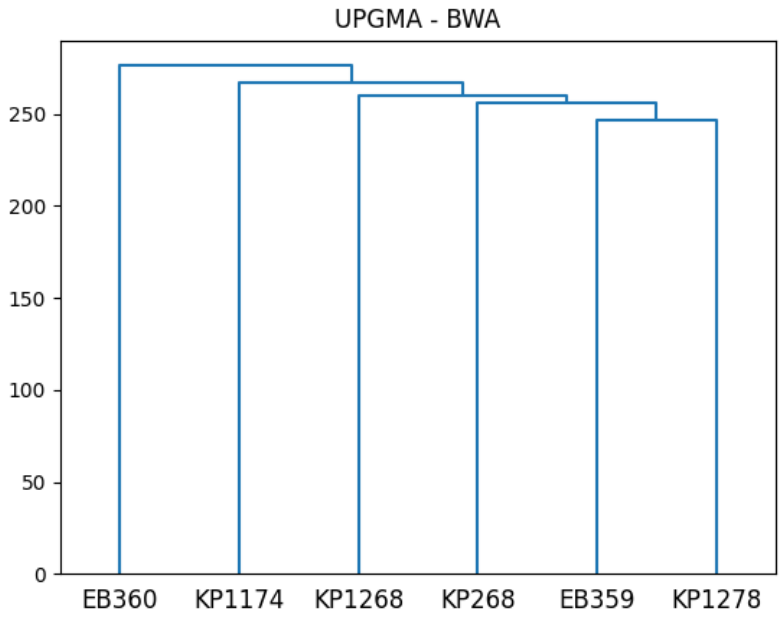
Obr. A.1: Minimální kostra grafu vzdáleností genomů sestavených assemblerem BWA



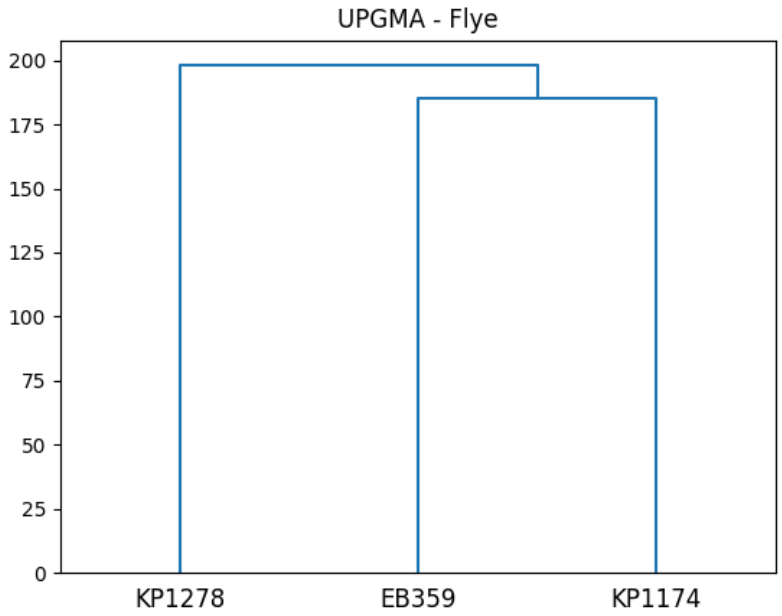
Obr. A.2: Minimální kostra grafu vzdáleností genomů sestavených assemblerem Flye



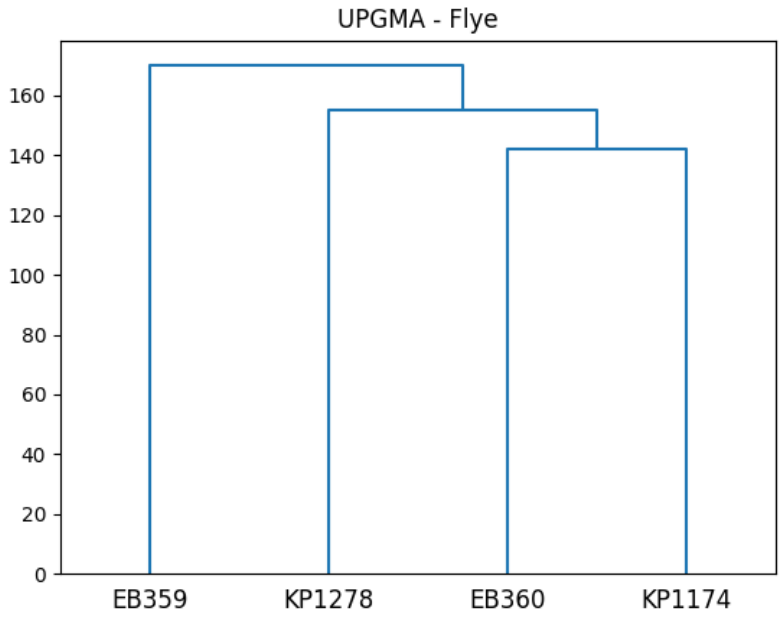
Obr. A.3: Minimální kostra grafu vzdáleností genomů sestavených assemblerem Flye



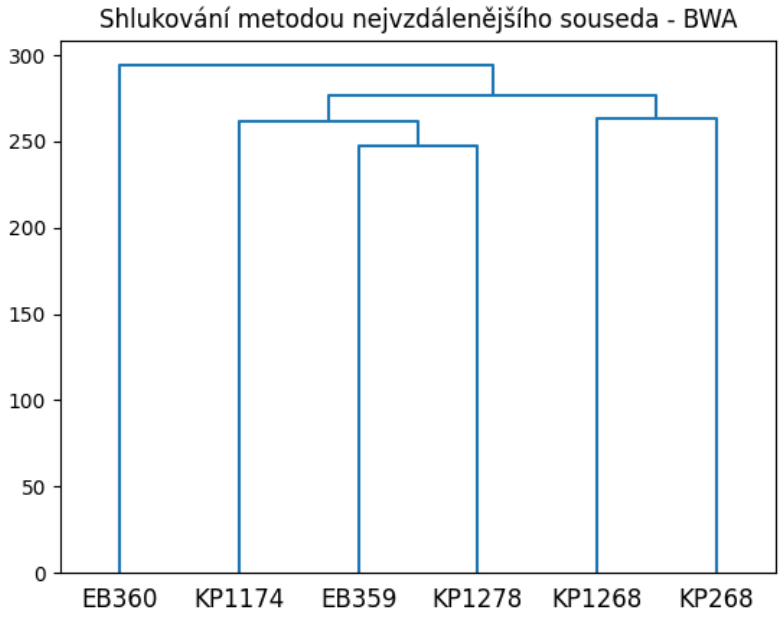
Obr. A.4: Dendrogram vytvořený metodou UPGMA z genomů sestavených assemblerem BWA



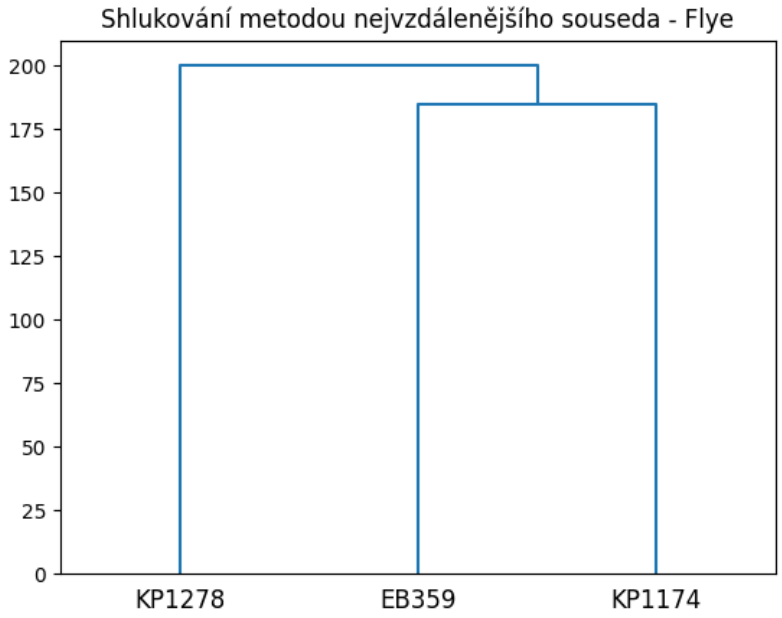
Obr. A.5: Dendrogram vytvořený metodou UPGMA z genomů sestavených assemblerem Flye



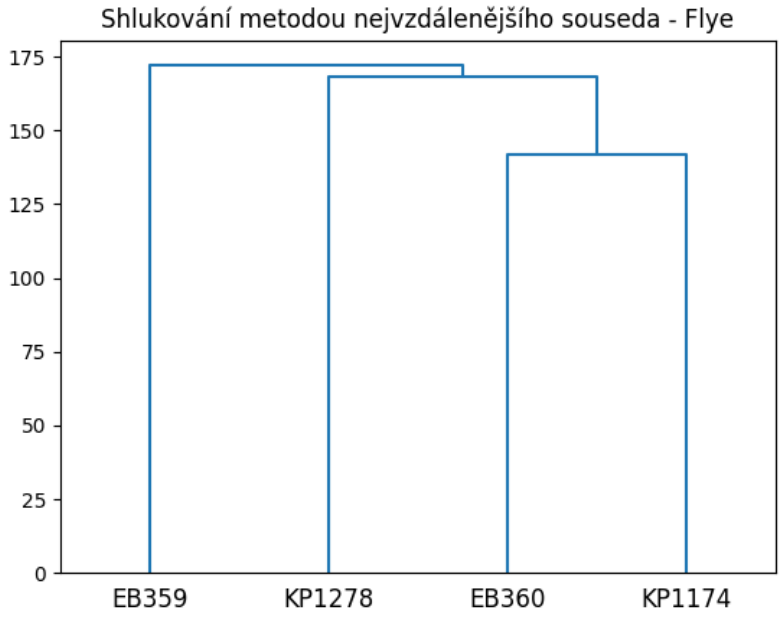
Obr. A.6: Dendrogram vytvořený metodou UPGMA z genomů sestavených assemblerem Flye



Obr. A.7: Dendrogram vytvořený metodou nejvzálenějšího souseda z genomů sestavených assemblerem BWA



Obr. A.8: Dendrogram vytvořený metodou nejvzálenějšího souseda z genomů sestavených assemblerem Flye



Obr. A.9: Dendrogram vytvořený metodou nejvzálenějšího souseda z genomů sestavených assemblerem Flye

B Obsah přiloženého ZIPu

V přiloženém zipu jsou jak skripty využité pro sestavování genomů, tak skripty využité pro následnou cgMLST analýzu.

```
/
├── BASECALLING_A_ASSEMBLY
│   ├── blast_master2.sh
│   ├── blast_slave.sh
│   ├── flye_extract_results.sh
│   ├── flye_master.sh
│   ├── flye_slave.sh
│   ├── guppy_alt.sh
│   ├── readme.txt
│   ├── spades_master1.sh
│   └── spades_slave1.sh
└── cgMLST
    ├── calculate_distance_matrix.py
    ├── clustering_and_plotting2.py
    ├── create_blast_database.sh
    ├── find_best_alleles.py
    ├── find_sufficient_quality_alleles.py
    └── make_mmld_table.py
```